DOCUMENT RESUME

ED 134 592 TM 005 513

AUTHOR Skakun, Ernest N.; And Others
TITLE The Use of a Rating Scale for Evaluating Performance in the Medical Field.
REPORT NO SME-76-2
PUB DATE Apr 76
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Certification; Evaluation Criteria; *Factor Analysis; *Factor Structure; *Medical Education; *Medical Students; *Rating Scales
IDENTIFIERS In Training Evaluation Report; Royal College of Physicians and Surgeons of Canada

ABSTRACT

The stability of the structure underlying a fourteen item rating scale (the In-Training Evaluation Report) when completed on two different groups of medical candidates was investigated at the Royal College of Physicians and Surgeons of Canada. In addition the relationship between the rating scale score and scores on other measures such as multiple choice and oral tests is reported. Results indicate that the factorial structures were similar for both groups of candidates and that the correlations between the scale and other measures were low. Implications for further research and improvement of the scale are also discussed. (Author/RC)

R.S. McLaughlin Examination and Research Centre

University of Alberta

Edmonton, Alberta

SME-76-2

THE USE OF A RATING SCALE FOR EVALUATING PERFORMANCE IN THE MEDICAL FIELD

Ernest N. Skakun
Donald R. Wilson, M.D.
William C. Taylor, M.B. Ch.B.

Presented at the Annual Meeting of the
American Educational Research Association
San Francisco

April, 1976

ABSTRACT

The stability of the structure underlying a fourteen item rating scale when completed on two different groups of medical candidates was investigated. In addition the relationship between the rating scale score and scores on other measures such as multiple choice and oral tests is reported. Results indicate that the factorial structures were similar for both groups of candidates and that the correlations between the scale and other measures were low. Implications for further research and improvement of the scale are also discussed.

The Use of a Rating Scale for Evaluating Performance in the Medical Field

Ernest N. Skakun, Donald R. Wilson and William C. Taylor

University of Alberta

In the area of medical education, evaluation, and assessment, a variety of procedures such as multiple-choice (MCQ), essay, and oral examinations have been used in the professional certification of medical candidates. More recently, rating scales have been used to assess specific dimensions of behavior such as problem solving ability, clinical judgment, and professional attitudes and global attributes such as physician performance and clinical competence (Barro, 1973; Linn, Arostegui, and Zeppa, 1975; Dowaliby and Andrew, 1976).

Starting in 1973, the Royal College of Physicians and Surgeons of Canada has used the In-Training Evaluation Report (ITER), a rating scale designed to provide additional information about a candidate's performance in delivering health care.

Since two groups of candidates (1973 and 1974) have been rated on the ITER, the purpose of the present paper is to investigate the factorial structure of the scale and to determine whether the emergent factorial structures are similar for the two groups of medical candidates. A further purpose is to determine the relationships between the ITER and other measures that are used in the certification process.

The In-Training Evaluation Report

The scale consists of fourteen criteria which have been derived from assessment sheets used in training programs in Canada and abroad and are believed to cover most of the important facets of a resident's profile while in training. The criteria used in the evaluation record grid describe a resident's overall performance on a day-to-day basis - his ability to assess patients, the quality of patient care, standards displayed in creation and maintenance of records and the resident's ability to function well as part of the health care team.

Each criterion is scored on a ten point numeric scale which is divided into five verbal gradations of unacceptable, poor, marginal, good, and excellent. To provide a common baseline for the raters, the end points of the scale for each criterion is defined in behavioral terms. That is, the attributes characterizing the unacceptable and excellent residents are listed for each of the fourteen criteria. Raters are thus provided with benchmarks which describe the qualities of candidates who should be rated as unacceptable or excellent. The middle three gradations of poor, marginal, and good are not defined in terms of candidate qualities. A sample of the criteria and their behavioral descriptions appear in Appendix A.

Method and Procedure

In 1973, a committee of not less than three members completed ITERs on 545 candidates who were registered and eligible for certification in one of the specialties of pediatrics, internal medicine, orthopedic surgery, urology, opthalmology, obstetrics and gynecology, diagnostic radiology or general surgery (Skakun, Wilson, Taylor and Langley, 1975). The same scale was completed on 778 candidates in 1974 who were likewise registered and eligible for certification in one of the above mentioned specialties. In

addition to ratings on the ITER, each of the candidates had a score on a multiple-choice examination and two oral tests. The first oral score was based on a clinical situation in which the candidate spent some time with a patient and then presented his findings to a pair of examiners. In the second oral, the candidate was questioned on hypothetical cases, management and treatment regimiu, and identification of aspects presented on slides, blood films, and radiographic material.

In an attempt to determine the dimensionality and structure of the scale, a correlation matrix of the intercorrelations of the 14 criteria was computed for each of the 1973 and 1974 data bases. To determine the factorial structure underlying the scale for the two administrations, the two correlation matrices were subjected to a principal components analysis. Using the criteria of eigenvalues greater than one and the scree test (Cattell, 1966) three components were extracted from each data set. The correlation matrices were then subjected to an image factor analysis followed by a normalized varimax rotation. Results of the rotated image factors based on the 1973 and 1974 candidates appear in Table 1.

Table 1 about here

The orthogonal procrustes method (Schönemann, 1966) was employed to assess the similarity of the two structures arising from the image factor analysis. Table 2 presents the results of rotating the structure based on the 1974 candidate to the target matrix resulting from the 1973 group.

Table 2 about here

Table 3 presents the intercorrelations between an overall rating on the ITER, the MCQ, Oral 1, and Oral 2. Values above the diagonal are for the 1973 data while those below the diagonal are for the year of 1974.

Table 3 about here

In order to provide an estimate of the internal consistency of the ITER, coefficient alpha, a generalization of the Kuder-Richardson 20 formula suitable in instances where the items are not scored dichotomously, was computed.

Results

In Table 1 if a factor loading falling outside the limits of $\pm.40$ is regarded as of practical importance, the first factor for the 1973 candidates is defined by criteria 1, 4, 5 and 6. The second dimension is defined by criteria 7 - 10 inclusive as well as criteria 2, 3 and 6. Criteria 11 to 14 of the scale define the last factor. Based on the results of the 1974 administration, factor I is defined by criteria 1, 3, 4 and 5 while the second factor is defined by criteria 2, 3 and 6 - 10 inclusive. The last dimension or factor is defined by criteria 11 to 14. It is evident from the factor analysis that the ITER scale is multi-dimensional in structure and that the underlying factorial structures emerging from the two sets of data are quite similar. However, factors I and II in both sets are not distinct as factor two is defined by criteria which also define factor one. Criteria 7 to 14 inclusive cluster according to the expectations of the design of the ITER scale and consistently for the two groups of candidates.

Since the results in Table 1 have arbitrary orientations within the three dimensional space which they define, it could be argued that the

numeric differences observed in the loadings between the 1973 and 1974 structures arose from differences in orientation and that an orthogonal rotation of one structure would closely resemble the other structure. This was investigated by rotating the 1974 matrix of loadings to the target matrix of 1973 using Schönemann's procedure (Table 2). The resulting least squares fit was good and interpretable, that is, the factors in Table 2 carry the same interpretation as that given to the two structures in Table 1.

The correlations between ITER and other measures are small suggesting that the linear relationships between these measures are almost non-existent.

Coefficient alpha can be interpreted as an index of the homogeneity of the criteria comprising the ITER scale. A coefficient of .82 was obtained for the 1973 data while one of .85 was obtained for the 1974 data.

Conclusions and Implications for Further Research

From the factor analysis it would appear that the underlying factorial structures are similar for the two groups of candidates. Both structures suggest that three separate aspects of candidate performance are assessed and that these might be tentatively labelled as Patient Assessment and Care, Professional Attitudes, and Technical Proficiency. There is some overlap between Factors I and II in terms of the criteria that define them and future work, therefore, should be directed towards purifying the first two factors. Criteria 2, 3 and 6 need some revision so as to align them more with criteria 1, 4 and 5.

It would appear that the ITER has a place in evaluating medical candidates by measuring attributes that are not assessed by the MCQ or the

Oral examinations. Obtaining sub-scores on the MCQ and orals as well as for the ITER and computing correlations and performing a factor analysis would enable a further interpretation of what attributes are assessed by the various tests.

At the present time the ITER constitutes an integral part of the final certifying process of the Royal College of Physicians and Surgeons of Canada. Because candidates are rated on the ITER by a committee of no less than three raters, the ITER no doubt suffers from the sources of error that are common to all numerical rating scales. Further research with the ITER should be conducted at improving the description of the criteria, defining more specifically what attributes of the candidate's capabilities the scale supposedly measures, and determining the effects of the various sources of error whenever rating scales are used. Studies to validate certain criteria of the ITER through the use of hospital charts and records and patient interviews are planned for the future.

TABLE 1

Rotated Image Factors Depicting the Clustering of the 14 Criteria of the ITER Scale for 1973 and 1974[1]

Criteria	Factors[2] (1973) I	II	III	Factors[2] (1974) I	II	III
Patient Assessment and Care						
1 History and Physical Examination	913	124	015	784	336	022
2 Clinical Judgment and Decision	217	674	266	377	680	151
3 Emergency Care	312	430	260	419	574	176
4 Comprehensive Continuing Care	855	131	007	793	256	-023
5 Laboratory Utilization	873	088	-006	813	267	-022
6 Records and Reports	417	425	118	387	641	117
Professional Attitudes						
7 Physician-Patient Relationships	093	799	153	196	741	200
8 Team Relationships	051	781	130	211	770	183
9 Ethics and Sense of Responsibility	118	815	163	196	749	197
10 Self-Assessment	056	704	187	178	699	222
Technical Proficiency						
11 Surgical Technique	-034	110	490	-013	040	401
12 Other Manual Skills Related to Specialty	082	183	626	009	169	736
13 Use of Equipment	-033	134	642	-013	162	732
14 Supervisory Skills	197	394	419	091	278	563

1 Factors rotated by the normalized varimax criterion.
2 Leading decimals omitted.

TABLE 2

Orthogonal Procustes Transformation of 1974

Factor Matrix to 1973 Factor Matrix[1]

Criteria	FACTORS I	II	III
Patient Assessment and Care			
1 History and Physical Examination	793	309	049
2 Clinical Judgment and Decision	393	672	142
3 Emergency Care	430	566	174
4 Comprehensive Continuing Care	803	227	008
5 Laboratory Utilization	822	237	009
6 Records and Reports	404	631	111
Professional Attitudes			
7 Physician-Patient Relationship	212	741	179
8 Team Relationships	229	759	162
9 Ethics and Sense of Responsibility	213	749	176
10 Self-Assessment	192	701	202
Technical Proficiency			
11 Surgical Technique	-032	057	397
12 Other Manual Skills Related to Specialty	-022	199	727
13 Use of Equipment	-044	193	723
14 Supervisory Skills	072	298	555

1 Leading decimal omitted.

TABLE 3

Pearson-Product Movement Correlations Between

ITER, MCQ and ORALS for 1973 and 1974[1]

	ITER	MCQ	ORAL 1	ORAL 2
ITER	-	.26	.04	.25
MCQ	.24	-	.08	.20
ORAL 1	.12	.22	-	.07
ORAL 2	.04	.20	.38	-

APPENDIX A

Two Criteria of the ITER

	Unacceptable 1 2	Poor 3 4	Fair 5 6	Good 7 8	Excellent 9 10	Not Applic.
A. Patient Assessment and Care						
1. History and Physcial Examination	_ _	_ _	_ _	_ _	_ _	___
B. Professional Attitudes						
1. Physician-Patient Relationships	_ _	_ _	_ _	_ _	_ _	___

Attributes of Residents for the Two Criteria

A. PATIENT ASSESSMENT AND CARE

A.1 History and Physical Examination

ATTRIBUTES OF A POOR RESIDENT	ATTRIBUTES OF A SUPERIOR RESIDENT
A poor resident takes incomplete and superficial histories which do not permit the development of good differential diagnoses. This type of resident displays an inability to ask the right questions and is disorganized in his ability to elicit information from the patient. Poor residents conduct incomplete physical examinations, miss important findings, or report abnormal findings which in fact do not exist.	A superior resident takes precise, reliable and comprehensive histories. Information is elicited in an organized and sequential manner which permits further investigation to proceed in a logical fashion. He displays good judgement in separating significant and insignificant patient statements. Superior residents carry out complete examinations and, where indicated, conduct detailed investigation of specific areas in order to make accurate diagnoses.

B. PROFESSIONAL ATTITUDES

B.1 Physician-Patient Relationships

A poor resident displays little compassion for or interest in the patient as a human being with medical problems and does not communicate well with the patient or his family. He is often critical of other members of the health care team in the patient's presence and does not inspire confidence in his patients. A poor resident displays little or no concern for patient morale.	A superior resident demonstrates a compassionate interest and an overall understanding of the patient as a person with an illness or an injury. He is patient and conscientious in explaining the nature of disease to the patient and his relatives and does not undermine the contribution of others. A superior resident inspires confidence in his patients and obtains their cooperation. At all levels he supports the morale of his patients.

REFERENCES

1. Barro, A.R. Survey and Evaluation of Approaches to Physician Performance Measurement. Journal of Medical Education, 48, 1051-1093, 1973.

2. Cattell, R.B. The Scree Test for the Number of Factors. Multiv. Behav. Res., 1; 245-276, 1966.

3. Dowaliby, F.S. and Andrew, B.J. Relationship between Clinical Competence Ratings and Examination Performance. Journal of Medical Education, 51, 181-188, 1976.

4. Linn, B.S., Arostegui, M., and Zeppa, R. Performance Rating Scale for Peer and Self-Assessment. British Journal of Medical Education, 9, 98-101, 1975.

5. Schönemann, P.H. A Generalized Solution of the Orthogonal Procrustes Problem. Psychometrika, 31, 1-10, 1966.

6. Skakun, E.N., Wilson, D.R., Taylor, W.C., and Langley, G.R. A Preliminary Examination of the In-Training Evaluation Report. Journal of Medical Education, 50; 817-819, 1975.